30th Annual International IEEE EMBS Conference
Vancouver, British Columbia, Canada, August 20-24, 2008

Leveraging Genetic Algorithm and Neural Network in Automated Protein Crystal Recognition

Ming Jack Po, Andrew F. Laine

Abstract—**We propose a classification framework combined with a multi-scale image processing method for recognizing protein crystals in high-throughput images. The main three points of the processing method are the multiple population genetic algorithm for region of interest detection, multi-scale Laplacian pyramid filters and histogram analysis techniques to find an effective feature vector. Using human (expert crystallographers) classified images as ground truth, the current experimental results gave 88% true positive and 99% true negative rates, resulting in an average true performance of ~93.5% validated on an image database which contained over 79,000 images.**

I. Introduction

PROTEIN structure determination is predominately solved through x-ray crystallography.[1] Unfortunately, there current exists no methodology to reliability predict crystallization conditions for a macromolecule that has previously not been crystallized. High-throughput experiments with varying crystallization conditions are currently performed with the hopes that one or more conditions will provide leads for actual protein crystallization. In a typical setup, each protein is mounted under thousands of conditions and crystallization is attempted for all the conditions simultaneously.[2] Consortiums in structural genomics such as Northeast Structural Genomics (NESG) now perform tens of millions of such micro-experiments annually, resulting in the need to analyze an even large number of images. Because crystals can form and dissolve in differing time scales, images from multiple time points are recorded per experimental condition. Unfortunately, since the images are currently classified manually, crystallographers can only inspect one time point per experimental setup in order to evaluate a specific experimental condition might have produced viable crystals.

Current proposed algorithms all involve the use of supervised learning algorithms. They mainly revolve around the use of neural nets [3-5] or the use of support vector machines [6, 7]. In current literature, both classes of supervised learning algorithms have been implemented without significant optimization and it is the quality of the training data and pre-processing steps that significantly alter the performance of the resulting classifiers. Unfortunately, none of the classifiers described in the literature thus far performs at fast enough speeds to be practical in a production structural biology pipeline. The current backlog of images at the NESG alone exceeds 50 million images. To finish processing this backlog in 5 years, an algorithm must process more than 20 images per minute as opposed to the current average speed of 30s per image [6].

We describe a classification framework that is being developed in collaboration with NESG to assist in the automated screening of protein crystal images. The three component of the classification algorithms are a genetic algorithm in order to determine the region of interest, a multi-scale Laplacian pyramid filter, and subsequent extraction of feature vectors used in a neural net classifier. Speed is of particular importance in our algorithm.

A. Pre-Processing

Due to the large number of images that are generated from high-throughput experiments, speed of execution is an extremely important consideration for any algorithm. Igor Jurisica's group from the Ontario Cancer Center have leveraged the World Community Grid, the world's largest public computing grid in order to tackle their image archives. [3]. We have chosen to take a two-prong approach, taking significant steps to optimize our algorithms for speed, while preparing to execute our program on the Google Computing Grid (current negotiations pending for Google App Engine).

Due to the variability of images that are captured by the robotics camera, finding the region of interest, an ellipsoidal droplet, using conventional algorithms such as the hough transform cannot be accomplished under a reasonable amount of time. As such, we have adapted a multi-population genetic algorithm in order to accurately locate our region of interest.

B. Network Classifier

After locating our region of interest, we then decompose the ROI with a multi-level Laplacian operator. Using the multi-level decomposition, we then compute different statistics that are used to train our classifier. Image classification is executed by a nonlinear feed forward neural network trained using mean square error optimization and back-projection.

The neural network itself is very easy to implement, and the key to the classifier's accuracy is the feature set that is chosen. With our current feature set, we calculate the mean, standard deviation, skewness, Kurtosis, energy, entrophy, area of enclosed regions, and linearity.

Manuscript received April 17th, 2008.

Ming Jack Po and Andrew Laine are with the Department of Biomedical Engineering, Columbia University, New York, NY 10027 USA. (Phone: 212-854-5996; fax: 212-854-5995; e-mail: ts2060@ Columbia.edu and laine@columbia.edu).

The classifier currently takes around a minute per image, and has accuracy around 90%.

II. Methodology

The flowchart of our method is shown in Fig. 1. The diagram shows the specific modules used to classify one image as well as the average execution time spent in module.

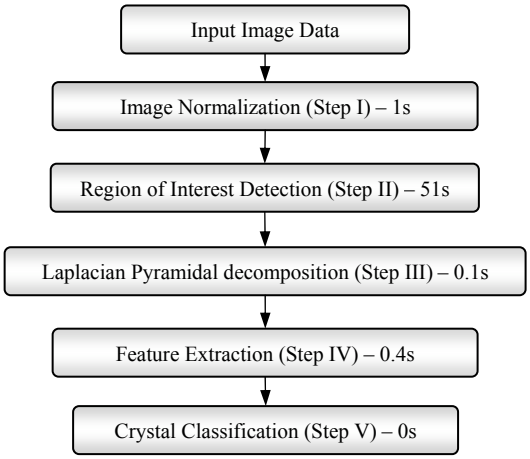


Fig. 1.Flowchart of integrated algorithm

A. Image Normalization and ROI Detection (Steps I-II)

Given a potential crystal image, we normalize the gray scale image by performing grayscale histogram equalization. Then, we construct an edge image using a canny edge detector. This edge image is then processed through an ellipsoidal multiple population genetic algorithm[8]. The details of the algorithms can be found in the reference above, but a brief overview will be provided here, along with specific adaptations that were implemented to better suit our purposes.

Chromosomes in the algorithm are no more than candidate ellipses. We know that each ellipse can be expressed in the standard equation

$$ax^2 + 2hxy + by^2 + 2gx + 2fy + 1 = 0 \quad (1)$$

From above, it is evident that each ellipse is uniquely determined through the parameters (a, h, b, g and f)[9]. The actual chromosomes store 5 different points, genes, on the perimeter of a candidate ellipse.

We begin the algorithm by randomly generating 100 potential candidates. The five points comprising each chromosomes are selected at random, from the foreground of the edge image. Then each candidate chromosome is evaluated based on two fitness criteria, similarity and distance. [4]. Essentially, the similarity score measures how close the five candidate points matches to an actual ellipse, while the distance score measures how far or close is the pattern to the ideal ellipse.

Evolution is carried out through both selection and diversification. Selection eliminates the ellipses that are particularly unfit, and replace those candidates with new candidates. Diversification allows for fit ellipses to pass onto the next generation through both crossovers and mutations.

The algorithm terminates when the convergence criteria is met. For our algorithm, convergence happens when no more subpopulations are created in 100 generations.

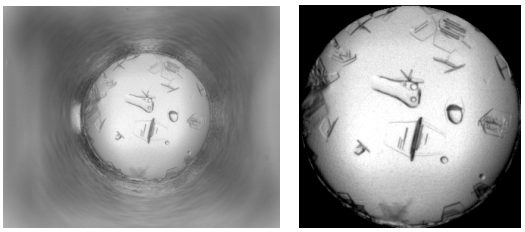


Fig. 2. Image Histogram equalization and ROI detection.

B. Laplacian Pyramidal decomposition and Feature Extraction (Step III - IV)

The Multi-scale Laplacian pyramid technique was originally used in image encoding [10]. Each level of the Laplacian pyramid is the prediction error L which is given by subtracting a low-pass filtered image from the original image. This Laplacian expansion is scale invariant, and thus allows us to compute global features without regards to orientation. The feature vector contains quantitative shape descriptions of the first, second and third-order histogram of Laplacian pyramid coefficients. This feature vector provides a more complete representation of the data driving the neural network as input.

The shape of an image histogram provides many clues as to the character of the protein crystal image. The selected quantitative shape descriptors of a first-order histogram are:

Mean: $$S_M \equiv \bar{b} = \sum_{b=0}^{L-1} bP(b)$$

Standard Deviation: $$S_D = \left[\sum_{b=0}^{L-1} (b-\bar{b})\right]^{1/2}$$

Skewness: $$S_S = \frac{1}{\sigma_b^3} \sum_{b=0}^{L-1} (b-\bar{b})^3 P(b)$$

Kurtosis: $$S_K = \frac{1}{\sigma_b^4} \sum_{b=0}^{L-1} (b-\bar{b})^4 P(b) - 3$$

Energy: $$S_N = \sum_{b=0}^{L-1} [P(b)]^2$$

Entropy: $$S_E = -\sum_{b=0}^{L-1} P(b) \log_2 [P(b)]$$

One of the second-order histogram features is used:

Autocorrelation: $S_A = \sum_{a=0}^{L-1} \sum_{b=0}^{L-1} abP(a,b)$

and

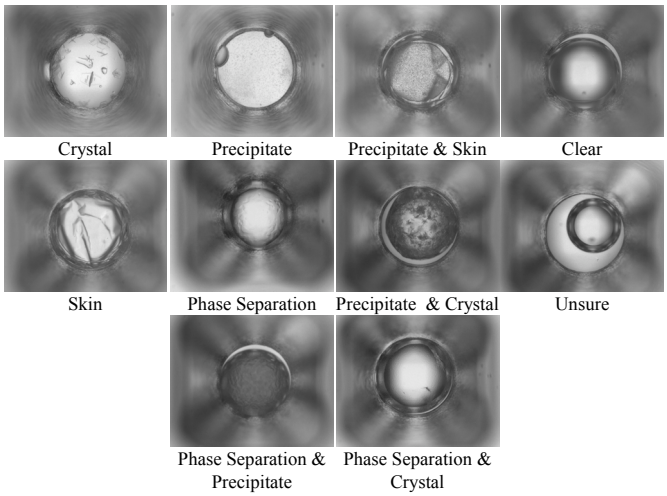
Power: $S_P = \sum^{L-1} FFT(\mathrm{Im})^2$

Where P(b) is the first-order histogram estimate. *P(a, b)* represents the histogram estimate of the second order distribution. Parameter *b* is the pixel amplitude value. *L* is the upper limit of the quantized amplitude level. Finally, σ is the standard deviation. *Im* refers to the image matrix.

In addition to these scale invariant features, we have added two more features that are only performed at the original resolution. These two features were added based on intuition from expert crystallographers. First, ellipses detected by a second run of the MPGA algorithm inside the ROI are extracted and their area summed. This is to better capture the fact that crystals are usually well defined enclosed shapes. We also used the linear Hough transform to complete a linearity score on the original resolution. This is to better capture needle crystals in images.

C. Expert Evaluation

The above procedure was tested on a data set of 79,632 classified images. These crystal images were each manually classified by 3 independent crystallographers at the Hauptman Woodward Medical Research Institute at the State University of New York at Buffalo. Each image was categorized into one or more of the following categories: Clear, Phase Separation, Precipitate, Skin, Crystal, Garbage, Unsure. For our purposes, we treated all images that were classified as crystals or crystals with additional categories by all 3 crystallographers as our ground truth crystal image. All other images were considered as non-crystal images. The 10 outcomes that arose are shown below.



III. Experiments and Results

Using the entire image database for training, and using leave-one-out cross validation, the current classification algorithm produced 88% true positive and 99% true negative rates on the validation dataset, resulting in an average true performance of ~93.5%).

The execution time on average is 12.5 seconds per image on a Core2Duo 2.4 Ghz machine with 6 GB of ram.

IV. Conclusions

With one focus on robustness of our classifier, our other focus was on speed of algorithms. Due to the enormous volume of images that high-throughput experiments are generating, a classifier cannot exceed 3 seconds an image without creating backlog. Since our classifier is a network based classifier, almost all of our processing time comes from pre-processing, i.e. the analysis of features from each image. The most time consuming steps have always been the identification of ROI in potential crystal images and the later calculation of feature vectors needed for either neural networks or support vector machines. With our implementation of the modified Multiple Population Genetic Algorithm, we have managed to decrease execution time to less than 15 seconds an image. Further work remains to be done to further decrease the processing time to less than 3 seconds per image without compromising the accuracy of the classifier.

Future refinements to the classification method include the use of more features, application of other pyramidal filters, and the further distinction of the microscopic images to separate drops with crystals, precipitates or organic matter.

We are also working with Google to prepare our current algorithm to run on their new Google App Engine. It is our hope that by leveraging the Google computing grid, we can further cut down our computational time from 60s an image to less than 20s an image.

V. ACKNOWLEDGEMENT

This project is part of the Northeast Structural Genomics Consortium (NESG) sponsored by the NIH for evaluating the feasibility, costs, economics of scale, and value of an infrastructure supporting high throughput structural genomics. The authors would also like to thank George DeTitta and the Hauptman-Woodward HTS lab for providing the labeled image data. We would also like to especially thank Angela Lauricella for her tireless hand-scoring of all the images for both the production NESG pipeline as well as for her instrumental role in the creation of the training dataset. We are also grateful to John Hunt and Gaetano Montelione for their expert advice on feature selection.

References

[1] Rhodes G, *Crystallography Made Crystal Clear*: Academic Press, 1993.

[2] C. R. Luft JR, Fehrman NA, Lauricella AM, Veatch CK, Detitta GT., "A deliberate approach to screening for initial crystallization

conditions of biological macromolecules.," *Journal of Structural Biology,* vol. 142, pp. 170-179, 2003.

[3] J. I. Cumbaa C, "Automatic classification and pattern discovery in high-throughput protein crystallization trails," *Journal of Structural and Functional Genomics,* pp. 195-202, 2005.

[4] L. S. Spraggon G, Kreusch A, and Priestle J, "Computational Analysis of crystallization trails," *Biological Crystallography,* pp. 1915-1923, 2002.

[5] C. C. Xu G, Angelini ED, Laine AF., "An incremental and optimized learning method for the automatic classification of protein crystal images.," in *Conf Proc IEEE Eng Med Biol Soc. 2006*, 2006, pp. 6526-9.

[6] S. G. Pan S, Penas-Centeno M, Xu D, Shapiro L, Ladner R, Riskin E, Hol W, Meldrum D, "Automated classification of protein crystallization images using support vector machines with scale-invariant texture and Gabor features," *Biological Crystallography,* pp. 271-279, 2006.

[7] T. M. Kawabata K, Saitoh K, Asama H, Mishima T, Sugahara M and Miyano M, "Evaluation of crystalline objects in crystallizing protein droplets based on line-segment information in greyscale images," *Biological Crystallography,* pp. 239-245, 2006.

[8] K. N. Yao J, Grogono P, "A multi-population genetic algorithm for robust and fast ellipse detection.," *Pattern Anal. Appl.,* vol. 8, pp. 149-162, 2005.

[9] I. J. Procter S, "A comparison of the randomized hough transform and a genetic algorithm for ellipse detection," in *Pattern recognition in practice IV: multiple paradigms, comparative studies and hybrid systems.*, K. L. Gelsema E, Ed.: Elsevier Science Ltd, pp. 449-460.

[10] Burt P.J, Adelson E.H., "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications,* vol. 31, pp. 532-540, 1983.